3C Empresa *(Edición núm. 25)* *Vol.5 – Nº 1*
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

Recepción: 30 de julio de 2015
Aceptación: 05 de febrero de 2016
Publicación: 22 de febrero de 2016

METODOLOGÍA HÍBRIDA PARA EL DISEÑO Y LA CONSTRUCCIÓN DEL DATA WAREHOUSE PARA "EL PROGRAMA DE REHABILITACIÓN AMBIENTAL Y SOCIAL" EN ECUADOR

HYBRID METHODOLOGY FOR DESIGN AND DATAWAREHOUSE CONSTRUCTION OF "EL PROGRAMA DE REHABILITACIÓN AMBIENTAL Y SOCIAL" IN ECUADOR

Ricardo Díaz Razo [1]

Oswaldo Díaz Rodríguez [2]

1. Egresado de Ingeniería de Sistemas e Informática, Universidad de las Fuerzas Armadas ESPE. E-Mail: ricardomdiazr@gmail.com
2. Master GIS, Escuela Politécnica Nacional. E-Mail: oswaldo.diaz@epn.edu.ec

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

RESUMEN

"El Programa de Rehabilitación Ambiental y Social (PRAS)" en el Ecuador almacena información hidrocarburífera y de gestión social del país en archivos semiestructurados y georreferenciados, lo que dificulta el cruce de información para el análisis y elaboración de reportes gerenciales y estadísticos. En el presente trabajo se propone una metodología para diseñar y construir un Data Warehouse con base en el punto de equilibrio entre los niveles de granularidad de la Organización y el grado de cohesión de sus funciones y con el pleno conocimiento de las mejores prácticas de las metodologías más conocidas en la industria de Data Warehousing (Inmon y Kimball); como herramienta integradora de datos y visualización se utilizó la plataforma "business intelligence Pentaho CE". La metodología se aplicó al PRAS y se consiguió reducir la complejidad en el diseño; además de la eficacia y escalabilidad en la generación de la información requerida.

ABSTRACT

"El Programa de Rehabilitación Ambiental y Social (PRAS)" in Ecuador stores information to hydrocarbon and social management of country in semi-structured and geo-referenced files, hindering the crossing of information for analysis and production of statistical and management reports. The present work proposes a methodology to design and build a Data Warehouse based on the balance between granularity levels of the Organization and the cohesiveness of its functions and with the full knowledge of best practices the methodologies more-known in the industry for Data Warehousing (Inmon and Kimball); as tool integrating data and visualization was used the platform "business intelligence Pentaho CE". The methodology was applied to PRAS and was managed to reduce design complexity, in addition of the efficiency and scalability in the generation of required information.

PALABRAS CLAVE

Punto de equilibrio; niveles de granularidad; Grado de cohesión; Kimball; Inmon.

KEY WORDS

Balance point; Granularity Level; Cohesion Degree; Kimball; Inmon.

3C Empresa (Edición núm. 25) Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

INTRODUCCIÓN

La Dirección de Investigación del PRAS orientada al procesamiento y sistematización de la información existente sobre daños históricos, demandas de actores afectados y notificaciones de los responsables, desarrolló en una primera etapa el Sistema de Indicadores de Pasivos Ambientales y Sociales dirigida a la actividad Hidrocarburífera Nacional (SIPAS-HN).

El PRAS actualmente posee gran cantidad de información sobre la actividad hidrocarburífera nacional en archivos planos y mapas que se encuentran sin estructurar ni depurar para ser procesada, esto se convierte en un gran inconveniente en el momento de realizar reportes o informes para la toma decisiones por parte de las autoridades. Es evidente la falta de un sistema de procesamiento electrónico de datos (transformación, conciliación, carga y análisis); por lo que se ha decidido implementar una plataforma de Business Intelligence Open Source (Data Warehuose).

En la actualidad, existen diferentes metodologías para la implementación de la solución, tales como Inmon y Kimball que permiten un enfoque descendente y ascendente (enfoque vertical) respectivamente a la vez que consideran datos atómicos y sumarizados, la metodología híbrida que se presenta en este trabajo contempla, además, un enfoque horizontal.

Del enfoque vertical descendente (Inmon, 2002) se obtuvo el último nivel jerárquico de cada dimensión de acuerdo con los niveles de granularidad de la estructura orgánica y funcional del PRAS y de acuerdo con el grado de cohesión (Kimball, 1996) definido por las reglas del negocio, se estableció el punto de equilibrio entre niveles de granularidad y grado de cohesión, lo que permitió generar un análisis transversal horizontal (Drill Across) entre dimensiones y medidas para crear un modelo analítico reutilizando las dimensiones comunes entre los Data Marts, esto ayudó a disminuir la complejidad de creación del modelo multidimensional y la escalabilidad en el momento de adicionar nuevos cubos de información al esquema (Díaz, 2015).

Como resultado de la creación de un modelo con estas consideraciones, se obtuvieron tres esquemas: Un esquema "public", en el que se encuentran todas las dimensiones utilizadas en el "Drill Across", un esquema "HN", en el que se encuentran las dimensiones únicas utilizadas por el módulo Hidrocarburífero Nacional y su tabla de hechos, y un esquema "GS" en el que se encuentran las dimensiones únicas utilizadas por el módulo de Gestión Social y su respectiva tabla de hechos.

Con base en los esquemas obtenidos se creó cubos de información con dimensiones compartidas, las que permitieron al usuario acceder a la información de manera transparente disponiendo de todas las dimensiones para sus consultas y al usuario técnico le facilitó la escalabilidad de creación de más modelos analíticos y cubos de información.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

METODOLOGÍA

Con base en las dos principales metodologías existentes se presenta el aporte de éste trabajo en la siguiente metodología, cuyo resultado de aplicación se muestra en el desarrollo de la solución para el PRAS.

Top-Down: Metodología de Bill Inmon que propone un modelo normalizado basado en la empresa, con una arquitectura de varios niveles y áreas de interés, Data Marts dependientes, poblando el Data Warehouse con datos a nivel atómico (Inmon, 2002).

Bottom-Up: Metodología de Ralph Kimball que propone un modelo dimensional de Data Marts, utilizando un esquema de estrella, con una arquitectura basada en las áreas de interés y Data Marts (Esparza, 2012), poblando el Data Warehouse con datos atómicos y sumarizados (Kimball, 1996).

Híbrido: Metodología propuesta por los autores de ese artículo, la que hace énfasis en el punto de equilibrio entre niveles de granularidad y grado de cohesión, con un diseño basado en el punto de equilibrio soportado por las reglas del negocio (Nagaoka, 2010), una arquitectura enfocada en el área del negocio, poblando los Data Marts con datos atómicos, sumarizados y concurrentemente compartidos.

Los modelos producto de los enfoques considerados (vertical ascendente, vertical descendente y transversal horizontal) no son diferentes, en el futuro mediato se verán similares y se complementarán en la aplicación.

MATERIALES Y HERRAMIENTAS

PostgresSQL: Sistema de base de datos relacional (diseñado para trabajar en conjunción con lenguajes de programación orientados a objetos como Java, C#, Visual Basic.NET y C++) que soporta distintos tipos de datos; además del soporte para los tipos de datos estándar, también soporta datos de tipo fecha, monetarios, elementos gráficos y geográficos, datos sobre redes, cadenas de bits, etc. Entre sus características más importantes se pueden mencionar, el sistema de seguridad mediante la gestión de usuarios, grupos de usuarios, permisos y contraseñas, gran capacidad de almacenamiento, licencia de tipo Berkeley Software Distribution (BSD), lo que permite manejar libremente su código fuente.

Pentaho: Esta herramienta se define a sí mismo como una plataforma de Business Intelligence (BI) *orientada a la solución y centrada en procesos* que incluyen los principales componentes requeridos para implementar soluciones cuyo fundamento son los procesos y se ha concebido desde el principio para estar basadas en procesos. Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes, integrado con un motor de workflow de procesos de negocio, la plataforma es capaz de ejecutar las reglas del negocio necesarias, expresadas en forma de un conjunto de actividades que entregan la información adecuada en el momento adecuado.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

DISEÑO E IMPLEMENTACIÓN

Se identificaron los siguientes macro procesos:

- Seguimiento, monitoreo y evaluación de la Infraestructura Hidrocarburífera y
- Ejecución de acciones previas y/o complementarias para la Gestión Social de la Rehabilitación Integral.

REQUERIMIENTOS DEL PROCESO DE NEGOCIO

En esta fase se identificaron los requerimientos del proceso de negocio, las necesidades de información de los usuarios que se agruparon en módulos para definir los indicadores correspondientes. Se determinó que es necesario partir desde un Enterprise Data Warehouse (EDW) para proceder a crear los Data Marts en esquemas estrella, para proporcionar a la institución una estructura escalable que permita incorporar nuevos indicadores en el futuro.

IDENTIFICACIÓN DEL GRANO (NIVELES DE GRANULARIDAD)

Identificando el grano se procedió a verificar que contenía un registro de una tabla de hechos exactamente. El grano muestra el nivel de detalle asociado a las medidas de las tablas de hechos. Fue necesario establecer el nivel de detalle que se desea estará disponible en el modelo dimensional.

IDENTIFICAR DEL GRADO DE COHESIÓN

Para la identificación del grado de cohesión de las funciones de la Organización se utilizó el método del Árbol de decisión que se muestra en la Figura 1.

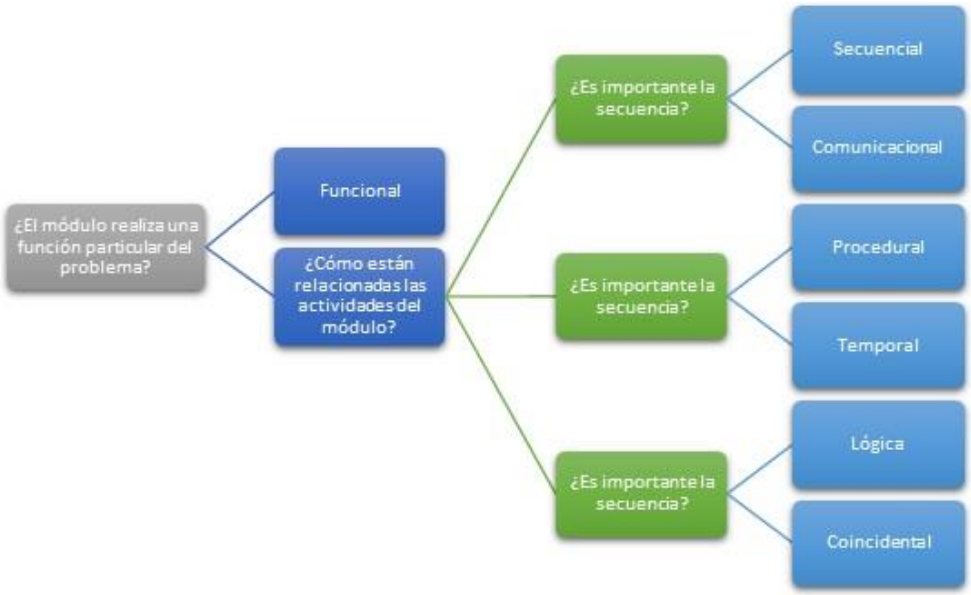

Figura 1: Árbol de decisión para el grado de cohesión. **Fuente:** Elaboración propia.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

IDENTIFICAR EL PUNTO DE EQUILIBRIO ENTRE NIVELES DE GRANULARIDAD Y GRADO DE COHESIÓN

Una vez identificados los niveles de granularidad y el grado de cohesión, se procedió a establecer un punto de equilibrio; para lo cual se analizaron todas las dimensiones una por una, considerando siempre las dimensiones en común.

INFRAESTRUCTURA	GESTIÓN SOCIAL	PUNTO DE EQUILIBRIO
Nombre	Detalle	
Sector Censal	Sector Censal	X
Micro Cuenca	Micro Cuenca	X
Área Protegida	Área Protegida	X
Bloque Petrolero	Bloque Petrolero	X
Campo Petrolero	Campo Petrolero	X
Territorio Indígena	Territorio Indígena	X
Estado Pozo	Tipo	
Tipo Estación	Actor Beneficiario	
Tipo Estatal	Tipo Documento	
Fecha	Fecha	X
	Ámbito Agravante Figura	

Tabla 1: Punto de equilibrio. **Fuente:** Elaboración propia

MODELO DE PUNTO DE EQUILIBRIO

El modelo de Punto de Equilibrio como se muestra en la Figura 2, se creó partiendo de los datos de la Tabla 1 entre los niveles de granularidad y grado de cohesión, aquellas columnas marcadas con X se convirtieron en características compartidas por los dos módulos, mientras que las que no eran comunes, se graficaron independientes en cada módulo.

Se colocó en el centro de la figura las dimensiones compartidas, en la parte superior las dimensiones únicas para el módulo de Infraestructura y en la parte inferior las dimensiones únicas para el módulo de Gestión Social.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

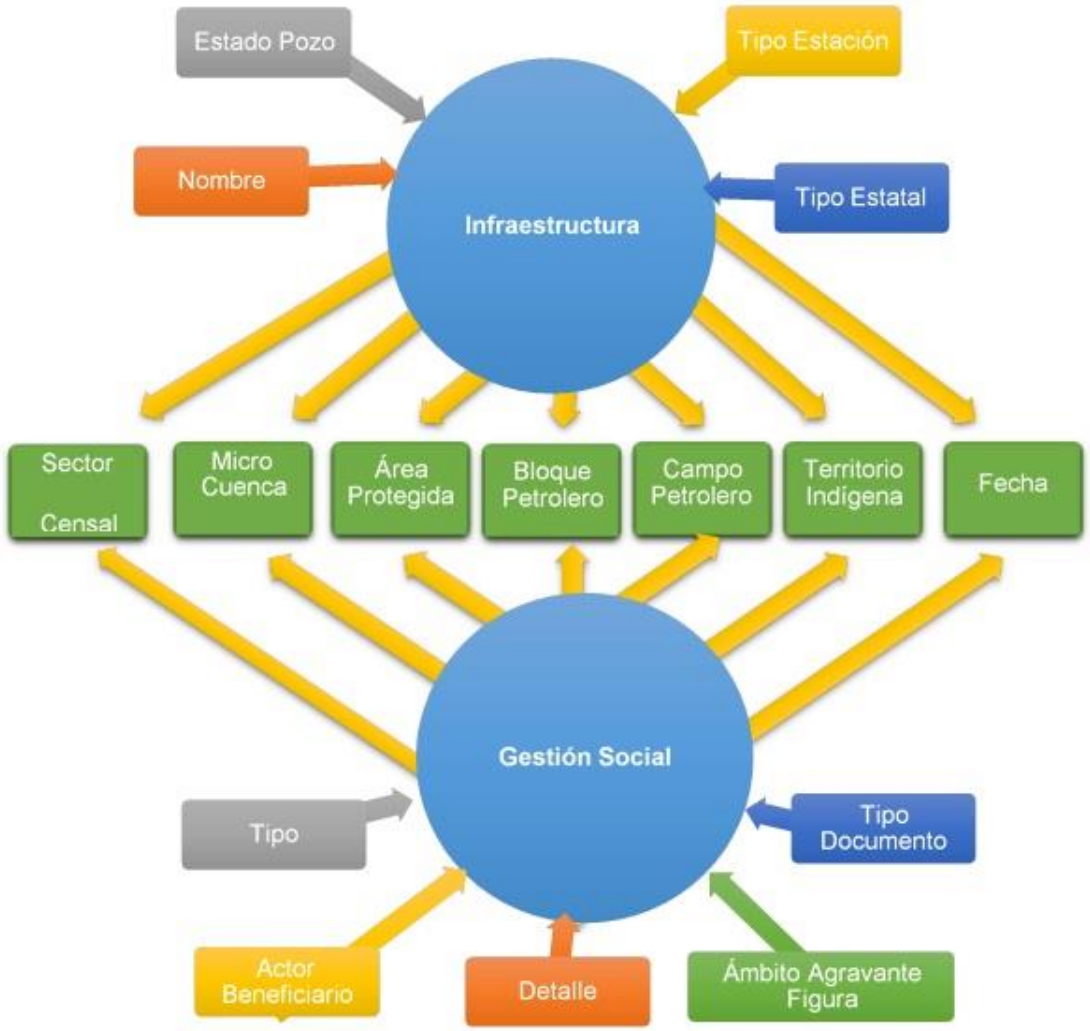

Figura 2: Modelo punto de equilibrio. **Fuente:** Elaboración propia.

IDENTIFICAR LAS DIMENSIONES Y MEDIDAS

Después de identificar el punto de equilibrio entre niveles de granularidad y grado de cohesión, se establecieron las tablas de dimensiones y medidas por cada tabla de hechos, así como las jerarquías de cada dimensión.

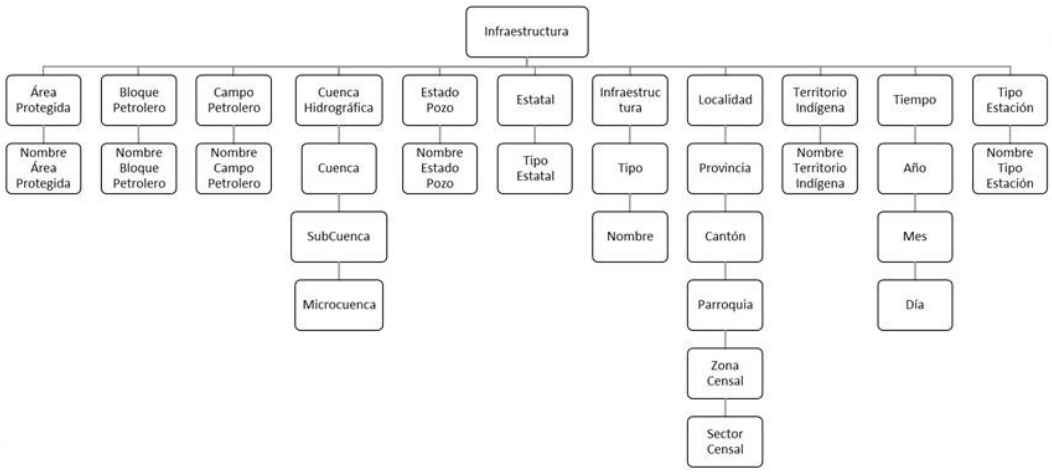

Figura 3: Drill Up y Drill Down infraestructura. **Fuente:** Elaboración propia

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

ANÁLISIS DE DRILL DOWN Y DRILL UP

De acuerdo al análisis de jerarquías de cada dimensión, se estableció el nivel máximo de detalle (Drill Down) del esquema, así como el máximo nivel de agrupación (Drill Up), como se muestra en la Figura 3.

IDENTIFICACIÓN DE DRILL ACROSS

De acuerdo al análisis previo, se determinaron las dimensiones en común entre las dos tablas de hechos (Figura 4), lo que permitió generar el Drill Across entre los esquemas considerando sus medidas.

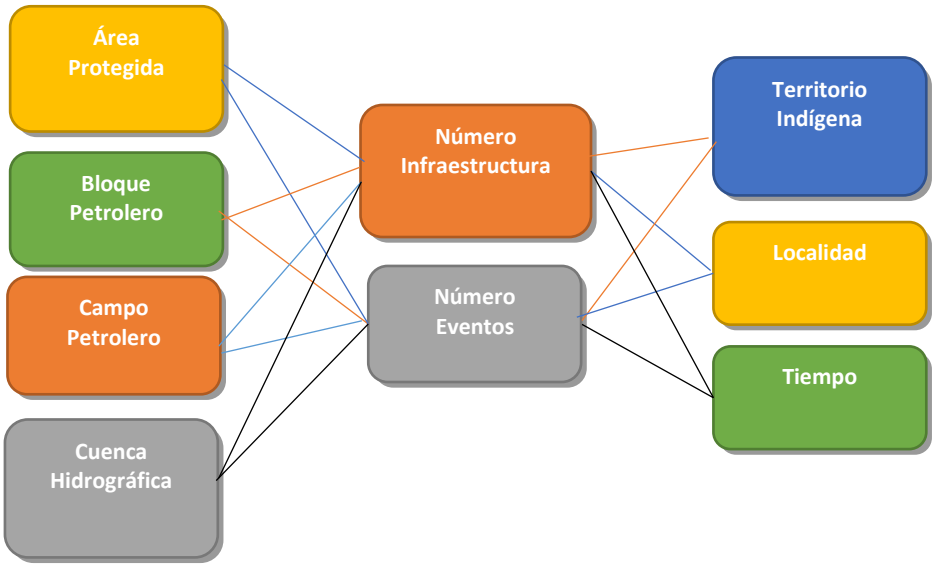

Figura 4: Drill Across. **Fuente:** Elaboración propia.

LIMPIEZA Y CALIDAD DE DATOS

Tablas de Staging: La información se extrajo desde ficheros semiestructurados y georreferenciados, se transformó su unidad de medida espacial de EPGS: 3260 a EPGS: 4326 para georreferenciar correctamente en OpenStreet Maps o Google Maps, de acuerdo con el proceso que se describe en la Figura 5.

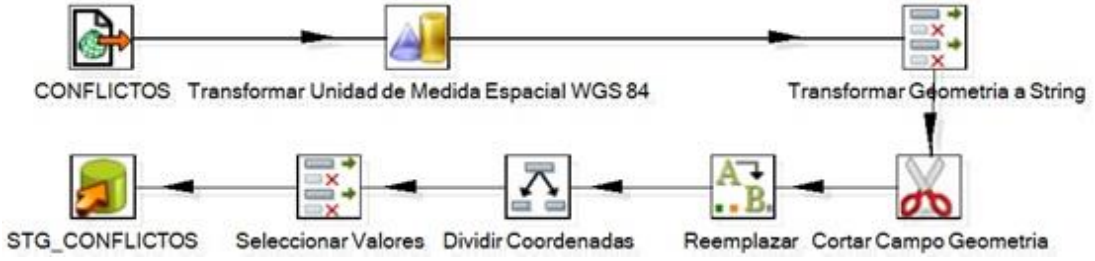

Figura 5: Extract, Transform, Load, tabla de staging. **Fuente:** Pentaho, elaboración propia.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

ALMACÉN DE DATOS (EDW)

A partir de las tablas de staging, se creó el modelo EDW, tal como lo describe Inmon (Inmon, 2002) y cuyo proceso se presenta en la Figura 6.

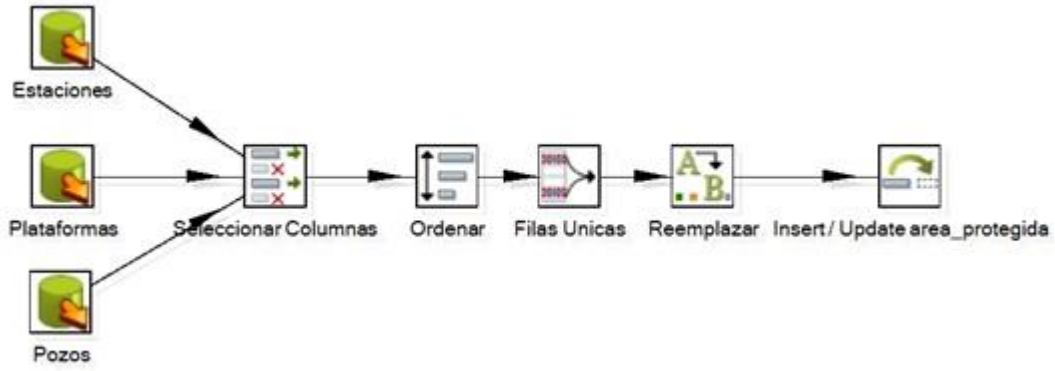

Figura 6: ETL EDW desde tablas de staging. **Fuente:** Pentaho, elaboración propia.

EDW

El modelo EDW se encuentra estructurado en tercera forma normal y tiene una estructura lógica relacional, considerando las mejores prácticas establecidas por Inmon (Inmon, 2002).

DATA MARTS

A partir del modelo EDW, se crean los Data Marts en estrella tal como lo describe Kimball (Kimball, 1996), teniendo en cuenta que las dimensiones compartidas se crearan una sola vez en un esquema público, mientras que las dimensiones que no sean compartidas serán creadas cada una en su esquema respectivo, de acuerdo con el proceso que se ilustra en la Figura 7.

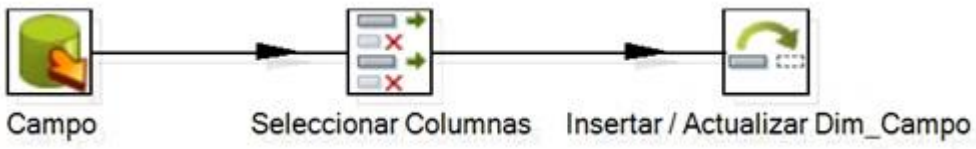

Figura 7: ETL dimensión Data Marts. **Fuente:** Pentaho, elaboración propia.

DATA MART INFRAESTRUTURA

Se verificó que el Data Mart tenía un diseño en estrella, considerando las buenas prácticas establecidas por Kimball (Kimball, 1996) para la creación de Data Marts.

3C Empresa ***(Edición núm. 25)*** *Vol.5 – Nº 1*
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

RESULTADOS

La estructuración de la información lograda que se muestra en la Figura 8, tanto del módulo de Gestión Social como del Módulo Hidrocarburífero a través del punto de equilibrio entre niveles de granularidad y grado de cohesión, reduce considerablemente el tiempo de creación del Data Warehouse ya que, a través de la identificación de dimensiones compartidas dentro del Drill Across, se puede realizar un diseño modular en el cuál cada dimensión constituye una parte única dentro de cada Data Mart exceptuando las dimensiones en común, las cuales son utilizadas para todo el Data Warehouse; estas no se encuentran creadas en el mismo esquema de cada Data Mart, sino que se encuentran creadas en un esquema público para que puedan ser accedidas independientemente.

Figura 8: Esquemas de la solución. **Fuente:** Pentaho, elaboración propia.

El resultado del modelo del cubo es representado en la Figura 9, donde  representa a las dimensiones únicas en cada cubo, mientras que  hace referencia a las dimensiones compartidas.

Figura 9: Cubo infraestructura. **Fuente:** Pentaho, elaboración propia.

3C Empresa ***(Edición núm. 25)*** *Vol.5 – Nº 1*
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

Desde la suite de inteligencia de negocios Pentaho, todo el modelo es *transparente* para el usuario final, para cualquier cubo de información que se consulte, se presentan las dimensiones necesarias para su análisis, gracias al diseño elaborado que se muestra en la Figura 8.

La Figura 10 muestra la evolución de pozos eliminados en la Provincia de Orellana desde el año 2000 hasta el año 2010, en el que se ve un pico de trabajos realizados por parte del Ministerio del Ambiente en el año 2005, en el que se eliminaron 120 pozos, mientras que para el año 2010 únicamente se ven reflejados 2 pozos eliminados; en este análisis se utilizan las dimensión compartidas *Tiempo* y *Localidad*.

La Figura 11 muestra el histórico del número de conflictos en la Provincia de Orellana desde el año 2000 hasta el año 2010, se puede ver un pico en el año 2007, en el que se presentaron 8 conflictos, en tanto que para el año 2010 únicamente se ven reflejados 2 conflictos; en este análisis se utilizan las mismas dimensiones compartidas que en la Figura 10, *Tiempo* y *Localidad*.

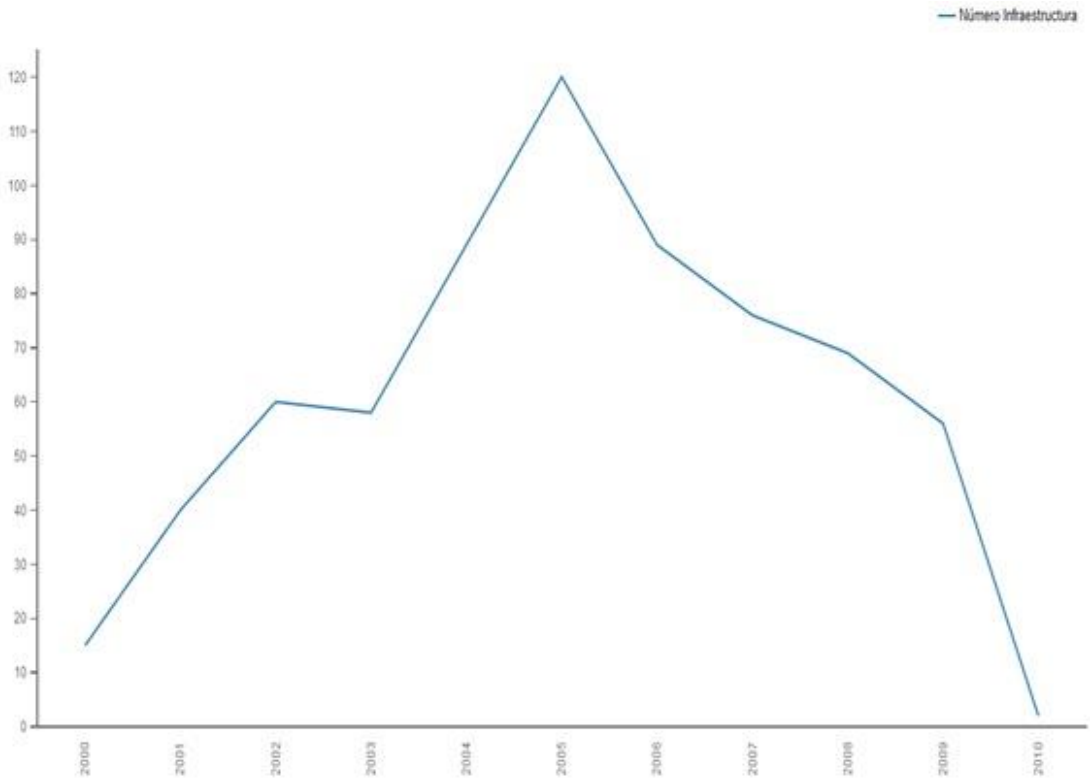

Figura 10: Evolución de pozos eliminados. **Fuente:** Pentaho, elaboración propia.

3C Empresa ***(Edición núm. 25)*** *Vol.5 – Nº 1*
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

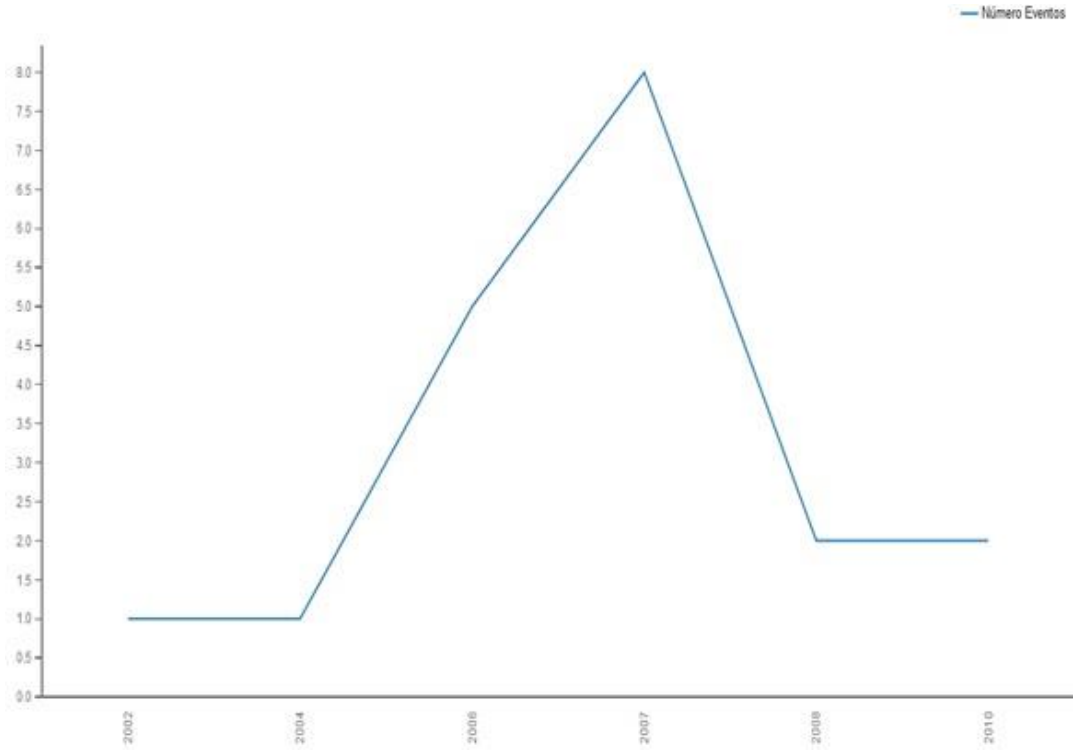

Figura 11: Histórico de conflictos. **Fuente:** Pentaho, elaboración propia.

Como se puede ver en los dos ejemplos, tanto la dimensión *Localidad* como la dimensión *Tiempo* forman parte de ambos cubos de información, sin embargo se encuentran creados una sola vez en el esquema *public*. El usuario desarrolló sus análisis sin inconvenientes y para él, toda la información se encuentra integrada en un solo lugar. Como valor agregado se crearon tableros de control para facilitar la visualización de información en pantalla, en la Figura 12 se puede ver el Dashboard de las Estaciones Hidrocarburiferas georreferenciadas en el país.

3C Empresa ***(Edición núm. 25)*** *Vol.5 – Nº 1*
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

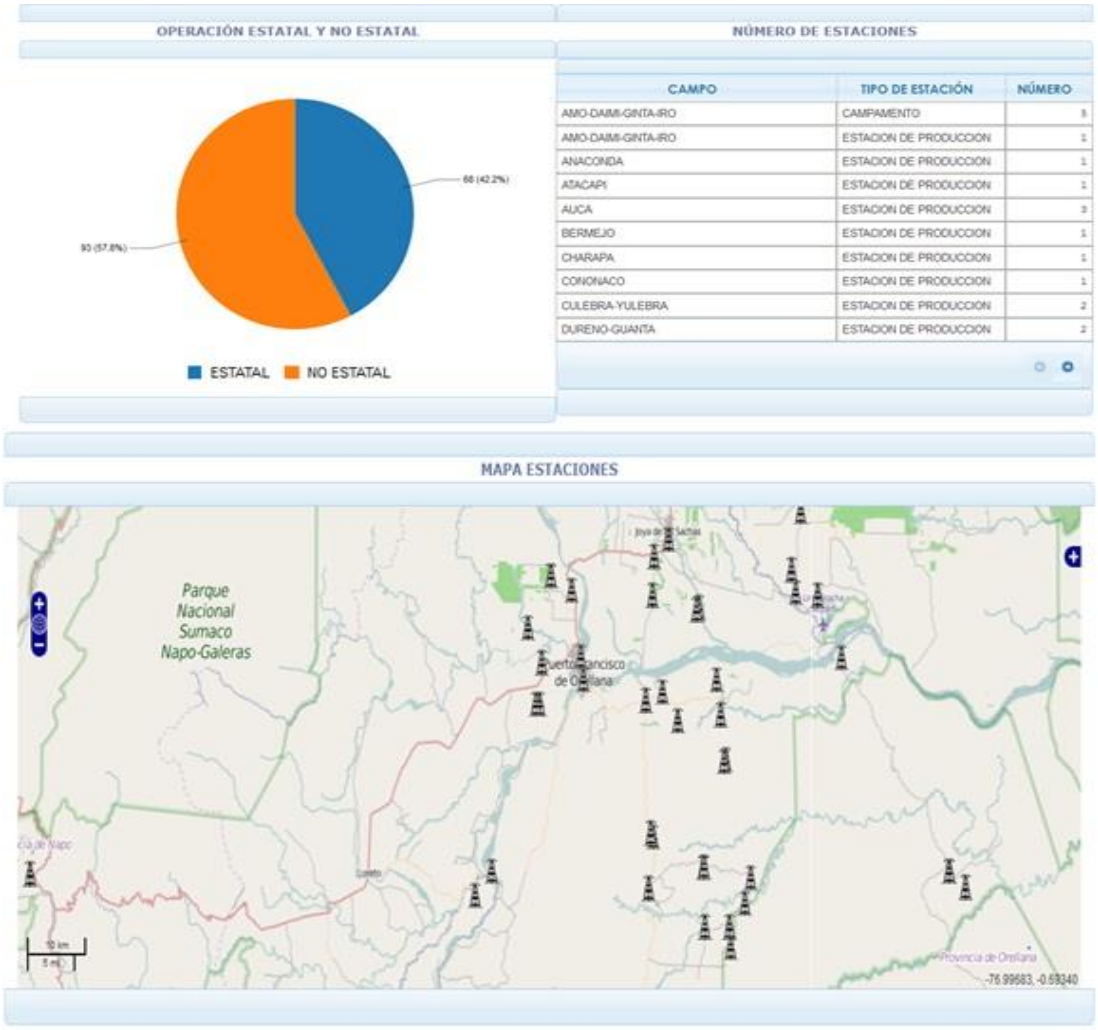

Figura 12: Tablero de control de estaciones. **Fuentes:** Pentaho, elaboración propia.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

CONCLUSIÓN

El pleno conocimiento de las metodologías de referencia (Inmon y Kimball), hábilmente combinadas y el dominio de las reglas del negocio llevan a la consecución de un robusto diseño dimensional y escalable en la implementación de soluciones Business Intelligence.

3C Empresa *(Edición núm. 25)* Vol.5 – Nº 1
Febrero – mayo '16, 1 – 15
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 3376
DOI: http://dx.doi.org/10.17993/3cemp.2016.050125.1-15

REFERENCIAS BIBLIOGRÁFICAS

Díaz Razo Ricardo. (2015). "Análisis y Estructuración de la Información Hidrocarburífera Nacional y Geoespacial Para el Diseño y Construcción de un Data Warehouse Para la Toma de Decisiones Socio-Ambientales del Programa de Reparación Ambiental Y Social - Pras". Biblioteca Universidad de las Fuerza Armadas ESPE.

Esparza, C.A. (2012). "Análisis, Diseño e Implementación de un Data Mart Utilizando Herramientas Open Source Para las Unidades Administrativa y Financiera De La Espe". Biblioteca Universidad de las Fuerza Armadas ESPE.

Inmon, W.H. (2002). "Building the Data Warehouse (Third Edition)". John Wiley & Sons, Inc. New York, NY, USA ©2002

Kimball, R. (1996). "The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data". John Wiley & Sons, Inc. New York, NY, USA ©1996

Nagaoka, H. Hitachi, Ltd. (2010). "Service Business Design Method Utilizing Business Dynamics" Service System and Service Management (ICSSSM), 2010 7th International Conference on 1-5.